By Lewis Johnson

For nearly a year now, ISD’s Tactical Language team has been working hard to develop training in less commonly taught languages and cultures. The Tactical Language Training System is designed to help people rapidly acquire basic communication and foreign language skills. It has two main components: a computer game called the Mission Practice Environment (MPE) and an intelligent tutoring system called the Mission Skill Builder (MSB). The Mission Practice Environment lets learners practice communicative skills through interaction with simulated people in a foreign culture. Learners choose gestures and speak phrases on behalf of their character in the game and see how the local people respond. The Mission Skill Builder helps learners to acquire the communicative and language skills necessary to accomplish their mission in the game. The MSB incorporates a virtual tutor that listens to learners speak, detects pronunciation errors, and gives learner tailored, motivational feedback.

When DARPA funded the Tactical Language Project, DARPA director Tony Tether issued a challenge: in order for us to obtain full funding for the project, we needed to get one of the military services to contribute funding. To do this, we would need to work very quickly to create a prototype training system that we could demonstrate to the services. So, building upon a commercial game engine (Unreal Tournament), we created an initial prototype within six months, and immediately proceeded to test it at West Point and show it to the services, particularly the US Special Operations training facility at Ft. Bragg, NC. Ft. Bragg is responsible for training Special Operations personnel (Continued on page 7)

By Hal Daumé III

If nothing else, one motivation that underlies nearly all work in our field of Artificial Intelligence is the desire to be able to generalize. Nearly all of the projects here in ISD fall under this rubric. Indeed, the explicit goal of the field of machine learning is to create systems that can predict attributes of unseen entities, based on generalized knowledge learned from similar entities.

Perhaps the strongest motivation for attempting to develop systems that generalize is that humans are so darn good at it. However, the level at which humans generalize is often quite different from the level at which machine learning and other AI programs attempt to generalize. In a canonical setting of multi-class classification, one would expect to be able to show a machine learning system a few pictures of lions, tigers and elephants, and then would expect the system to be able to predict, upon seeing a new lion, tiger or elephant, to which of the three groups it belongs.

Certainly, humans are able to perform this sort of generalization readily. However, humans can often go one step further than the stereotypical machine learning system. Most humans (even small children), in seeing these groups of lions, tigers and elephants, will be able to not only distinguish among them, but also to distinguish new animals. For instance, having seen only lions, tigers and elephants, a machine learning system might classify a giraffe as a lion (for its color) or as an elephant (for its size), but a human will immediately recognize that this is a previously unseen type of animal. Furthermore, if a child is then shown another giraffe, he
May 2004

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14 Joe’s Birthday!</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>17</td>
<td>Marc’s Birthday!</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>24</td>
<td>dg.o 2004</td>
<td>25</td>
<td>dg.o 2004</td>
<td>26</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>31</td>
<td>Memorial Day</td>
<td>ISI CLOSED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

June 2004

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>Lewis’ Birthday!</td>
<td>9</td>
<td>Emil’s Birthday!</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>Ann’s Birthday!</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>Mei’s Birthday!</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>27</td>
<td>Nick’s Birthday!</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SCHEDULE OF EVENTS:**

**May:**
- May 2-7: HLT/NAACL 2004
- May 3: Cristyn Beenau’s Birthday!
- May 8: Jafar Adibi’s Birthday!
- May 14: Joe Chiu’s Birthday!
- May 17: Marc Spraragen’s Birthday!
- May 21: Brett Rutland’s Birthday!
- May 21: ISD Poster Fair
- May 24-26: dg.o 2004
- May 31: Memorial Day — ISI CLOSED!

**June:**
- June 8: Lewis Johnson’s Birthday!
- June 9: Emil Ettelaie’s Birthday!
- June 15: Ann Myer’s Birthday!
- June 20: Mei Si’s Birthday!
- June 25: Jose-Luis Ambite’s Birthday!
- June 28: Nick Mote’s Birthday!
**ISD Tidbits**

- **Janice Cole** left the division in February and moved back to her hometown in Michigan. **Ann Myers** will be taking her place as our new Financial Business Manager (the position formerly known as Administrative Coordinator). Ann brings much experience to ISI from closely related work in private industry, and she is quickly coming up to speed on the university environment. Stop by and welcome her to the division when you have a chance!

- In this year’s TREC competition in , the system entered by **Soo-Min Kim**, **Deepak Ravichandran**, and their advisor, **Ed Hovy**, to identify opinion-bearing sentences and event sentences in texts, scored second out of 35 competitors from around the world. Contratulations!

- **Craig Knoblock** was recently elected as a AAAI Fellow! Congrats on this extremely high honor!

- **Brent Lance, Lei Qu, Nick Mote and Marc Spraragen** were recently screened into the Ph.D. program at USC. Congratulations to all!

- **Shou-de Lin** was chosen to present his Ph.D. thesis work, "Interesting Instance Discovery in Multi-Relational Data," in the AAAI 2004 Doctoral Consortium, in San Jose, with a $1000 scholarship award.

- **Franz Och** has left ISI, and **Dimitra Papachristou** will be leaving on May 7. They are heading to Google and will be missed. You can now reach Franz at och@google.com. Dimitra hasn’t been assigned an email at Google yet, but she will keep us posted. Good luck Franz and Dimitra!

- From now on, The InSiDer will be featuring one article per issue from an ISD student. If you are a student and would like to write about your research, please contact Kathy Kurinsky at kurinsky@isi.edu. If you are not a student, please be on the look-out for the student articles to find out what students in the division working on!

- During ISD’s retreat this year, the division decided to hold regular ISD poster fairs. The first one will be held on May 21 from 10:30 am—12:00 pm in the 11th floor conference room. All are welcome and encouraged to participate (especially students!). Please contact Lei Ding at leiding@isi.edu if you are interested in participating. You can also find more information at http://brahms.isi.edu:8080/twiki/bin/view/ISD/PosterFair.
DONOVAN ARTZ
DONO@ISI.EDU

I am a new Ph.D. student, and my primary research interest is in representing and reasoning about complex ideas. My current work at ISI is focused on understanding and possibly extending existing work on the Trellis project, addressing trust relationships between users and sources. I am privileged to be working with Yolanda Gil and the Interactive Knowledge Capture group at ISI.

I am most recently from Philadelphia, where I attended Drexel University, receiving both my Bachelor's and Master's degrees in Computer Science. My previous research includes Steganography, Machine Learning, Mobile Ad hoc Networks, Information Assurance, Security, and Autonomous Agents. My former affiliations include the Geometric and Intelligent Computation Laboratory at Drexel, and the Advanced Weapons Design group at Los Alamos National Laboratory.

CRISTYN BEENAU
CRISTYN@ISI.EDU

Hello! I want to say thank you to everyone for making me feel so welcome here at ISI! You all have probably seen me walking through the halls at one time or another. A little bit about me: I am a temporary assistant, employed by AdvectA, who has been helping out the admins for about two months now, and I plan on being here for at least another four months. I am currently a student at Santa Monica College, majoring in Computer Information Systems. Thanks again for making me feel so welcome!

HAI HUANG
HUANGH@ISI.EDU

Hello everyone! I am Hai Huang, currently in my final stage of the M.S. Program in Computational Linguistics. I am working with Kevin Knight on some machine translation projects. My other research interests include computational linguistics, web mining, information retrieval and CJK (Chinese, Japanese and Korean) Information Processing. I was born in Shanghai, China and lived in more than a dozen different places before I was ten, which stimulated my initial interest in language. I have spent a great deal of time at universities, trying to figure out which career path best suits me. Nevertheless, I did have some real world experiences at two great companies. Before I came to the U.S., I worked at IBM's sales office in Beijing for a year, where I helped maintain their production scheduling database. Last year, I worked on the search engine implementation team at eBay, in San Jose. My hobbies include collecting antique books, hiking, travel, soccer, classical music and Jazz, following current events and learning languages. My aspiration has been to become either an entrepreneur or an author.

SALIM KHAN
SKHAN@ISI.EDU

My name's Salim, and I am a Ph.D. student with Dr. Craig Knoblock's Infoagents group. I came here from the University of Delaware, where I completed my Master's degree. Originally though, I am from Bangalore, India—yes, that Bangalore. I am interested in bioinformatics-related data integration, and when I am not integrating data sets, I like to play cricket, follow politics, go clubbing and read non-fiction.

CHARLIE LAC
LAC@ISI.EDU

Hi, my name is Charlie, and I'm currently working alongside Stacy Marsella on the MRE project. This is a collaboration between ISI and ICT. My role on the project is the procedural animation or the way that characters move with complex artificial intelligence.

CHIRAG MERCHANT
CMERCH@ISI.EDU

I was born in 1979, in Mumbai (Bombay), India, and I am currently pursuing my Master's in Computer Sci-
Chirag Merchant
Room 452

Ann Myers
Room 925

Josh Moody
Room 944

Thiago Pardo
Room 960

ence at USC, specializing in Multimedia and Creative Technologies. My areas of interest include graphics, animation, special effects and web technologies. Someday I hope to render actors useless.

I joined ISI in January, 2004 and am currently working on the Tactical Language Training Project in the CARTE Lab, with Lewis Johnson and Hannes Vilhjalmsson. Before coming to the USA in 2002, I worked as a software engineer in India and Japan.

I love cartoons, movies, TV and music. Popeye is my role model 😊 For more blabber about me, http://www.isi.edu/~cmerchan is the place.

ANN MYERS
AMYERS@ISI.EDU

I was born in Tokyo, Japan, moved first to Ohio for a few years and then on to New Mexico where I was raised. After attending the University of New Mexico, I moved to California. My experience is in office management with a financial and statistical analysis background gained at Princess Cruises, Haworth Inc., BKM Inc. and DPM, Inc. While at DPM, I worked with the Department of Homeland Securities, regarding sales, proposals and contracts for the White House, the Pentagon and numerous DHS locations throughout the US. My position with ISI is Financial Business Manager, reporting to Yigal Arens.

I love reading, hiking, biking, camping and spending time with a good friends who love to attend outdoor concerts, movies and group dinners.

JOSH MOODY
MOODY@ISI.EDU

I am one of the members of Paul Cohen's group, who moved here last fall from the University of Massachusetts Amherst. There, I was completing my undergraduate degree in Computer Systems Engineering. I am now enrolled at USC in the Computer Science department. At ISI, I am working for CRUE (Center for Research on Unexpected Events). Currently I am focusing on the AIID (Architecture for Interpretation of Intelligence Data) architecture, which merges a blackboard system with incrementally built Bayesian networks. Most of my adult life, I worked as an itinerant shepherd in Maine, New Hampshire, Vermont and Florida for several different land management programs. I am an avid cyclist, commuting to ISI from downtown Los Angeles. I’d like to thank everyone at ISI for making me feel welcome.

THIAGO PARDO
THIAGO@ISI.EDU

Hello, my name is Thiago Alexandre Saguio Pardo, and I come from Brazil. I was born in Bauru, a city in the middle of the state of São Paulo. As a curiosity, the word bauru is an Indian word and means “basket of fruits.” It is also the name of a very famous sandwich in Brazil, made of ham and cheese. You can imagine all the jokes about this.

I arrived at ISI in the beginning of the year and intend to stay here until November, working on my Ph.D. research about discourse parsing, under Daniel Marcu’s supervision. In fact, I am in the third year of my Ph.D. courses at the Universidade de São Paulo, in Brazil and am supposed to finish them next year. In my Master’s research, I worked on text generation, summary planning and extractive summarization. Previously, I also worked on the development of a grammar checker for Brazilian Portuguese and investigated several grammatical formalisms in a comparative study.

In my free time, I like to be with my friends, play volleyball (not necessarily well...indeed, ignore the word “necessarily,” to be more exact), ride my bicycle, watch movies and read Marvel comics and science fiction and horror books, especially those by Stephen King. My two particular passions are X-Men and The Lord of the Rings. I’m sorry to disappoint those who thought I could play soccer and dance samba, as most Brazilians supposedly can. I just can’t do these (weird) things, but I do appreciate watching those who can.


By Deepak Ravichandran

I arrived in Los Angeles in the fall of 2000 with a little more than $1,000 in my pocket (thanks to my dad!), a Bachelor’s diploma, an admission to the USC Ph.D. program (thanks to Ed Hovy!), and a large heart. I knew ISI was far from the USC main campus, but what I didn’t know was that the public transportation here was very poor. After thinking carefully about it, I decided to stay near USC, since I would be taking a lot classes on campus and since I didn’t have any money to buy a car. I spent the first few days taking care of registrations and other formalities, and I soon became friends with two other new Indian students.

The next day, one of them declared that we should all go to the Santa Monica beach. After all, USC is not exactly the kind of place you think of when you imagine Los Angeles—especially after seeing all of those Hollywood movies. We three brave young souls embarked on the mission of going to Santa Monica by bus. It was a bright and sunny August Saturday afternoon. First, we had to take a bus downtown for the connection to Santa Monica. We got there, and after half an hour wait, we finally boarded the bus to Santa Monica. The bus inched along through the surface street traffic, and after some time, a friend of mine pointed towards the window at the majestic “HOLLYWOOD” sign. “We sure are in Los Angeles,” I declared, feeling elated.

After some time, a large man boarded and sat down in front of us. He was about six-and-a-half feet tall, had very long hair, and was decked out in a cap and dark glasses. For some reason, I thought that he was acting weirdly, but it was not until a few moments later that it became clear to me that this man was totally drunk. I was hoping that he wouldn’t start any trouble—but I was wrong. Apparently we looked as strange to him as he did to us. “Where you guys from?” he demanded. We all ignored him initially. Convinced that he wouldn’t start any trouble, we continued on our way.

So that they can work effectively in a variety of foreign countries. Over the subsequent months we incrementally improved and extended the system.

In March we demonstrated the system at DARPATech in Anaheim. A variety of people came by our booth, including decision makers at Special Operations and members of the press. Both came away very impressed. Special Operations Command drafted a Memorandum of Agreement with DARPA to co-sponsor our project. Slate Magazine and National Public Radio’s Day to Day mentioned our work as one of the most interesting projects at the conference.

Now that the agreement has been signed, we have more work ahead of us: we must put the first version of our training system into use this fall at Ft. Bragg. This is a great challenge as well as a great opportunity. We need to show that the system really works— that it enables more people to learn basic foreign language skills more rapidly than conventional approaches do. The underlying technologies (learner modeling, socially intelligent tutorial dialog, multi-agent pedagogical drama, learner speech processing, automated collaborative authoring, natural language processing, focus of attention tracking, etc.) must be reliable, and the overall system must be able to recover from component errors and failures. We must employ state-of-the-art assessment methodologies, so that we can show conclusively that the system is effective and show how the component technologies contribute to the system’s effectiveness. And last but not least, the game should be fun—learners should be motivated to play the game and practice their communication skills on their own. We have a lot to do, but we are getting strong support from DARPA and Special Operations for our work. And if we succeed, we will be well on our way to changing the way foreign languages are taught. Check back with us this fall and see how well we did. And in the mean time, come try the game yourself— you will find yourself conversing in Arabic in no time!
Where in ISD???
Make yourself at home

ISDers will go to great lengths to make their tiny office space abodes more like a “home away from home.” Beanbag chairs, a change of clothes, not-so-subtle reminders to take it easy... you’d be amazed at what you can find in your colleagues offices! Can you remember where you’ve seen the items below? Stumped? The answers are on page 14.

Object 1
Marina Towers

Object 2

Object 3
relax more at work

Object 4

My First Few Days in America, continued

(Continued from page 7)

we were avoiding him, his tone grew louder. "Where the hell are you guys from?" he shouted. One of my more courageous friends decided that there was no point in avoiding him. He replied, "We are from Los Angeles." That seemed to irritate the drunk stranger even more. He kept on, "I am not asking about that. Where are you guys originally from?"

That got my friend thinking. We wanted to give an answer that would not irritate this guy, but it was possible that he didn't like people from certain places. So my friend replied with a very good answer. "We're from Asia." The stranger exploded in laughter, "That is so funny, because you look like people from that country where Saddam Hussein came from!" (Remember, this incident took place only a year before 9/11, so I was getting really scared.) This strange foreign country was freaking me out!

The bus ride seemed to go on forever. I looked at my watch; it was already 35 minutes since we boarded the bus, and the ocean was nowhere in sight. We had been interrogated by a very scary looking guy—what was next? Just then, the bus made a scheduled stop, and a very young attractive female boarded. The drunk stranger (in mid-conversation) suddenly lost his concentration. He diverted his attention to the young woman, deciding that his afternoon would be better spent with her than with us. He followed her to the rear of the bus and sat next to her. We were relieved—no more uncomfortable interviews by strange scary looking people. But after five minutes, the pretty woman started shouting, asking the drunk to mind his own business, and pleading with the driver for help. The driver pulled the bus to the curb and ordered this guy to get off of the bus. After some aggressive persuasion by the driver and some threats to call the police, the man finally got off the bus and left us to ride in peace. The rest of the day was relatively normal. We visited Santa Monica (the bus ride was about two hours each way). Because this incident happened the first week that I was in Los Angeles, I decided never to go to Santa Monica by a bus again.

A week after the above incident, I was traveling to ISI from USC (by bus again). The ride is about an hour and a half with a connection at the intersection of Figueroa and Slauson Street. For people who are not familiar with Los Angeles geography, this is the heart of South Central Los Angeles: 90049, the zip code with the largest amount of violence per year in Los Angeles. It was a Tuesday, September morning. It was hot. I looked at my watch; it was about 10:30 a.m., and it looked like I had just missed my bus connection. I checked the time-table. Sure enough, the next bus wasn't coming for 50 minutes. I was stranded!

I stood there as cars piled on the on-ramp of the Harbor 110 freeway, and after a while, I saw someone who also seemed to have missed the bus. I decided to engage him in conversation (Remember this was my second week in Los Angeles, and I was a lot braver then). He seemed like a nice guy in his late twenties. He told me about how he was attending Cal State Los Angeles and was majoring in Psychology, about how he was working his way through college and had had a very tough life. He even talked about the racial problems in America that he had encountered. Things seemed normal enough. After some time into the conversation he asked me the normal question, "Where are you from?" and I replied back, "I am from India." He quickly said, "Oh India! Isn't that near the East coast?" I started thinking that maybe for this man the entire world revolved around Los Angeles and that anything not near Los Angeles must be near the East coast. I was wondering whether I should explain that India is a far-off land on the other side of the globe. But in the end, I decided that I was not in the mood to lead a geography lesson. I smiled back replying, "Yes, India is near the East coast." The bus came an hour later, and I eventually reached ISI.

Living near USC and traveling to ISI by a bus was always adventurous. One of my close friends was mugged on the way home from having dinner at my apartment. Once a robber slipped into our apartment when no one was there and decided that there was nothing worth stealing, except for my roommate's toolbox. Another time, someone stole my neighbor's car but left my car unharmed (I was elated because I did not have theft insurance). I always met an amazing number of weird people on the bus—at least enough to keep me entertained during the journey. But there is something to be said for my early Los Angeles adventures. Now that I have my own car and have moved into a neighborhood that more resembles "Suburbia Americana" things are very boring and are never as exciting as they used to be.
The InSiDer

Thanks to a grant from the Bavarian State Government, CARTE has initiated a collaborative project with the Lab for Multimedia Concepts and Applications at the University of Augsburg. The Lab for Multimedia Concepts and Applications, headed by Elisabeth André, is one of the leading European research laboratories in multimodal human-computer communication. They develop animated agents capable of multimodal interaction in augmented reality environments and other enriched communication environments, and work with a range of input devices including vision processing, speech recognition and analysis, and biological sensors.

Some areas of particular common interest between CARTE and the University of Augsburg are social communication and emotion modeling. Both groups, for example, have applied the sociolinguistic theories of Brown and Levinson to the problem of modeling dialog. The Lab for Multimedia Concepts and Applications has examined both verbal and non-verbal aspects of social communication, and has developed a model conversational agent that can vary its communication tactics based on the emotional state of the user.

The collaborative project between CARTE and the Lab for Multimedia Concepts and Applications will permit us to undertake cross-cultural studies of social communication. One such study will involve taking the dialog interaction tactics developed in CARTE’s Social Intelligence Project, recasting them in German, and evaluating their perceived politeness in the German educational context. Other projects will be initiated and developed through reciprocal visits between the two laboratories.

More Thoughts for Food

By Yolanda Gil

After reading Deepak’s inspiring “Thoughts for Food” piece in the last InSiDer, I thought that I would share my own choices of eateries around Marina del Rey and the South Bay. Deepak’s “Calories per dollar” is not the metric here; rather, I rate the food quality and entertainment value per dollar. Enjoy!

Killer Shrimp (523 Washington Blvd., Marina del Rey, 578-2293) offers an interesting concept: a restaurant that only serves shrimp. Three shrimp dishes done with the same simmering spicy sauce. They also offer an interesting concept in terms of their waiting staff that I’ll let you discover for yourself.

The best Japanese restaurant in LA, according to one of ISI’s most food-savvy Japanese visitors (and of yours truly, for what it’s worth), is a stroll away from ISI. East (4317 Glencoe Ave., Marina del Rey, 822-3700) has a great range of appetizers and sushi options. Once you find out about their interesting table arrangements, everything will taste even better.

For a true home-cooked, down-South, hearty meal, you want to go to Aunt Kizzy’s Back Porch (4325 Glencoe Ave., Marina del Rey, 578-1005, www.auntkizzys.com). Fried chicken, collard greens and corn bread are some of the main staples. It’s about time you learned that hush puppies aren’t necessarily shoes! All absolutely delicious, so you definitely want to avoid their all-you-can-eat Sunday brunch!

When you can’t get a hold of East coast farm coffee, treat yourself to Joni’s Coffee Roaster and Juice Joint (548 Washington Blvd., Marina del Rey, 305-7147). They roast their own beans every morning at the crack of dawn. Warning: you may never want to set foot in the local S_ _ _ _ _ _ _ again…and if so, good for you!

For those of us who cannot get excited enough about Thai restaurants, there is Chan Darette (13490 Max- ella Ave., Marina del Rey, 301-1004, www.chandara.com). Their vegetable curries are the best. To make your palate rock, ask for the coconut rice or saffron rice.

If you like restaurant adventures or you are a real meat-lover (or both), you must try Samba (207 N. Harbor Drive, Redondo Beach, 374-3411). All-you-can-eat Brazilian churrasco is their specialty, with waiters bringing new meat choices from the grill to your table every few minutes and with a great salad bar of side dishes and soups.

Craving great Chinese food but not willing to drive deep into the inland empire? W’s Chinese Bistro’s (1410 Pacific Coast Highway, Redondo Beach, 792-1600) owner prides himself on using the freshest products and highest quality ingredients, and it shows. And how many Chinese restaurants do you know that have a dessert chef?

You will be magically transported to Louisiana if you go to the Ragin’ Cajun (422 Pier Avenue, Hermosa Beach, 376-7878). Browse through their wait-bench library before you feast on gumbo, jambalaya, blackened catfish or (just flown-in from the bayou) crawfish etouffé. Twenty hot sauces on every table, a Dixie and Cajun music will take you miles away.

If you ever dreamed of brunch at the beach, Martha’s (23 22nd Street, Hermosa Beach, 376-7786) is hands-down the best of the best. They know how to do classics like eggs benedict the right way, but you won’t regret being adventurous with their signature white corn scramble, tabouleh scramble or Thai scramble. And for those of you who like a mystery as well as good food, I will tell you that Martha’s has a very curious connection to ISI…
will almost certainly be able to identify it as a giraffe (even if he does not know the name of this new animal). Seeing then a hippo, a child should once again be able to identify it as a new sort of animal.

While the above example of zoology might seem contrived, it is actually the case that such problems bear their heads in many contexts. For instance, one might wish to be able to classify documents or books or emails by the original author— to do so, one must be able to generalize in the way described above, as it would be infeasible to enumerate and obtain training data for every possible author. In another context, one of the central problems in computer vision is to be able to track the motion of objects in a movie; however, before tracking can take place, one must be able to identify which parts of an image correspond to the same object. Or, one might wish to be able to merge two databases that use different names for similar fields.

My particular interest in this problem arose in the context of the coreference resolution task. In this task, one is supposed to take a document and identify all noun phrases that refer to the same entity. For instance, consider the very short document: "George Bush gave a speech today. The President said his advisors were very helpful." In this document, there are five noun phrases (also called "mentions"); they are: "George Bush," "a speech," "The President," "his," and "advisors." The job of a coreference system is to take these mentions and report that "George Bush" and "The President" and "his" are all mentions of the same entity, while "a speech" and "advisors" are both singleton mentions (i.e., they refer to unique entities).

This problem is just like our earlier discussion of wild animals. We have at our disposal a few hundred documents that are annotated for both mentions and entities. These annotations are the equivalent of telling a child about lions, tigers and elephants. The system is then presented with a new document, with completely new mentions and entities (and old ones too), and is asked to identify them, just as the child is expected to be able to identify giraffes and hippos.

A typical machine learning solution to this problem (and to others with the same characteristics) is to attempt to make a learning system that distinguishes between two classes: same entity or not same entity. This binary classification system is trained on pairs of mentions from the same document and then applied to pairs of mentions in a new document. There are many problems with this solution, both theoretically and practically.

The biggest problem, which spans the space of theory and practice, is that unless this classifier is perfect (and it is, in general, very far from perfect), it will often make the mistake of producing incompatible predictions. For instance, in our above example, it might decide that "George Bush" and "The President" are coreferent (or refer to one another) and that "The President" and "his" are coreferent, but that "George Bush" and "his" are not. Many heuristics, including greedy search, and clustering methods, have been proposed to get around this problem. However, such solutions achieve far from optimal performance.

Another problem is that of complexity. Most documents contain several hundred mentions; if we examined all pairs of mentions individually, there would be literally hundreds of thousands of examples to consider. Moreover, the vast majority of these examples are "not coreferent," leading to what is known as the imbalanced class problem: machine learning systems, in general, tend not to perform well when one class vastly outnumbers the other.

From a theoretical standpoint, proofs of optimality for nearly all machine learning systems require that the training examples that they learn from are independently and identically distributed. In simple terms, this means that knowing the correct answer to one of the yes/no questions should give us no information about the answers to any of the other questions. By training a system on such pairs of instances, it is clear that this assumption is horribly violated: the training example comparing "George Bush" to "The President" is, in fact, quite dependent on the training example comparing "George Bush" to "speech," by virtue of the fact that they have an element in common. While it is unclear whether this issue makes a difference in practice, it is something of which we need to be aware.

For the purpose of pure aesthetics, as well as for all of these reasons, I have decided to build a system that attempts to solve the original problem, without reducing it first to binary classification. The current approach that I am entertaining is based on recent techniques from Bayesian statistics.

From a Bayesian perspective, we can describe this process in a generative fashion: we treat the mentions and their corresponding entity assignments as a collection of data and ask ourselves how this data came into being—the "generative story." The story we adopt is the following: we assume that there is an infinite collec-

(Continued on page 13)
The creation of any new research discipline is difficult, but the creation of a Digital Government research community is particularly so. This is due to the broadness of the problems investigated, the absolute need for interdisciplinary collaboration, and the requirement for support from government agencies that are often accustomed to involvement in research.

Outreach and community building activities are therefore essential for the budding Digital Government community. This grant is the third in a series we have received in support of such activities.

The grant will support the following:

1. Organizing the annual National Conference on Digital Government Research. This is a venue for publishing, showcasing and sharing the best of digital government research results. This year it will be held May 23-26 in Seattle.

2. Informational activities: A web site for information about digital government projects, software to support collaboration among different parties, and a newsletter and news clipping service for the digital government community to keep it informed of events and new opportunities.

The truth about real-world learning is surely more complicated than "procedures first, facts and reasons later;" yet this very rough theory has some attractive features. First, one can do useful work with procedures, even if one doesn’t completely understand them. This means learning is grounded in activity and is gradual over the lifetime of an agent; learning isn’t a one-time application of a batch algorithm to a dataset. Second, exercising procedures produces the context and the failures one needs to learn non-procedural knowledge, such as facts and reasons that we call understanding. Third, non-procedural knowledge is learned not in its most general form, but in the form required for completing a procedure. Subsequent executions of the procedure provide the cases one needs to generalize non-procedural knowledge. Because generalization is grounded in procedure execution, overgeneralization is less likely. Fourth, because real-world learning takes place in a social environment, and procedures unfold step by step over time, there often is a human to help the learner correct mis-steps before a procedure goes completely wrong. This ensures a second type of gradualism, what we might call within-task gradual learning, and minimizes the credit assignment problem, in which a learner must figure out which of a long sequence of things it did wrong.

In sum, learning by doing is how agents extend their knowledge—procedural and factual—over their lifetimes while doing useful work, and so is ideal for Enduring Personal Cognitive Agents (EPCAs). Learning by Doing wraps naked procedures in the cloth of domain knowledge, reasons and justifications, so the EPCA, while always capable of doing useful work, gradually comes to understand what it is doing.

The overall goal of this project is to develop novel statistical and linguistic techniques that will exploit the information that is available in parallel multilingual corpora (i.e., translations of the same source in multiple languages). Such corpora implicitly encode a hidden, common core that can be uncovered using state-of-the-art parameter estimation techniques.

The research plan involves two main thrusts: i) automatic learning of structure in and across languages at multiple levels of abstraction: seman-
The training data that we have is only particular referral type. "advisors" are generated by two (the first by full name, the second by generated using different referral types. They are, however, presumably generated from the same underlying entity. "his" in the above example are generated.

For instance, all three mentions, "George Bush," "The President," and "his" in the above example are generated from the same underlying entity. They are, however, presumably generated using different referral types (the first by full name, the second by title, and the third by a pronoun). The other two mentions, "a speech" and "advisors" are generated by two other, separate entities, with some particular referral type.

The training data that we have is only annotated for mentions and groups of coreference, which means that the actual entities and the referral types are unknown, "hidden" variables. Since, in general, the entities appearing in the training data will be different from the entities appearing in the testing data, we don't particularly care about what the entities are, but rather only how they relate to each other. We can thus view the referral types as telling us how groups of co-referring entities interact. The task of learning then boils down to inferring what these referral types are, and how many there are. Once these are known, when we present an unseen document to the system, it will be able to use the knowledge it learned in the form of referral types to uncover the underlying entity sets.

While such a model is conceptually very attractive, the math required to deal with the potentially infinite sets of entities and referral types quickly becomes quite hairy. One is forced to make certain simplifying assumptions about the form of the models used, in order to be able to perform inference in such a model. The trade-off is between specifying models that are sufficiently simple as to admit tractable inference, and specifying models that are sufficiently motivated as to make the results meaningful. My current research is in investigating how to make these decisions.

While my primary focus is on the coreference task, I believe that such algorithms will be more widely applicable than this constrained task. In the end, the result of such work will be a system that can learn to generalize from knowns (such as lions, tigers and elephants) to unknowns (giraffes and hippos), hopefully moving the generalization capability of machines closer to that of children.